This version: 21 Dec. 2010 Dr. John Bendix

Peer Judgment vs. The Indicators

A case study of *Anglistik* in Germany

Introduction

The advocates of peer review argue that only those in a discipline, preferably those with adequate experience of and in it, can properly judge the worth or value of its products and programs. That qualitative argument is countered by those who advocate assessment based on quantitative indicators. Some analysts (Henk Moed; Hans-Dieter Daniel) argue for a middle ground of "informed peer review," but there appear to be few cases where quantitative indicators and qualitative judgment are applied to the same objects. A published CHE survey of *Anglistik* programs conducted in 2007 across Germany, however, allows one to compare the two.

An initial overview

In 2007, the CHE examined *Anglistik* at 52 German universities using three indicators: (A; see Table 1) the amount of outside or third party funding spent (*verausgabten Drittmittel*); (B) "the results of a bibliometric analysis," and (C) the number of doctoral degrees (*Promotionen*) awarded. These indicators were used to construct a list of "research strong" programs – which was given prominence by being placed in a summary table at the beginning of the report – that were assigned points (D).

Tucked away at the end of the report, and explicitly declared to be informational, was a ranking of "leading" programs by reputation (E). Interestingly, while this reputational ranking did not form a part of the numerical, point-based, assessment of "research strong" programs, if that program had a reputation as "leading," the information was presented alongside the summary table. So while the quantitative indicators were given prominence, the qualitative judgment of reputation was regarded as supporting evidence, where applicable. The three indicators are divided so as to be able to identify which programs belong to a top (T), middle (M), and bottom (B) group. This differentiation corresponded to the cut-off points, through cumulative addition in descending order, of a ranked distribution. Thus, the top 50% of the distribution (T) was marked with a green dot, the next 40% (M) with a yellow dot, and the bottom 10% (B) with a red dot. That these colors correspond to those of a traffic

light is not accidental, and has much to do with the purpose of providing prospective students with a guide to which programs are more – and less – desirable, at least on these indicators.

One can gain an initial impression from Table 1:

Table 1

Anglistik Programs Listed

(in descending order; program = bottom of T group)

	A Funding 49 programs	B Publications 53 programs	C Degrees granted 51 programs	D Research-strong 10 programs			E Reputation 12 programs
1.	LMU München	Giessen	LMU München	LMU	(6)	++	LMU
2.	Giessen	LMU München	Freiburg	Giessen	(6)	++	Freiburg
3.	Mannheim	Freiburg	Heidelberg	Freiburg	(4)	++	Giessen
4.	Erlangen	Bamberg	Duisburg-Essen	Münster	(4)	+	FU Berlin
5.	Münster	Mainz	Bochum	Regensburg	(4)	+	HU Berlin
6.	Marburg	Regensburg	Hamburg	Bayreuth	(3)	++	Tübingen
7.	Kassel	Frankfurt	Giessen	FU Berlin	(3)	++	Frankfurt
8.	Bayreuth	Münster	Regensburg	Heidelberg	(3)	++	Mainz
9.	Freiburg	Heidelberg	Würzburg	Konstanz	(3)	+	Hamburg
10.	HU Berlin	Leipzig	Mainz	Mainz	(3)	+	Regensburg
11.	FU Berlin	Köln	Frankfurt	++ = research strong			Heidelberg
12.	Düsseldorf	Jena	Münster	also in 2004			Bremen
13.	Bremen	Konstanz	Trier	+ = newly research			Köln
14.	Augsburg	Bonn	RWTH Aachen	strong in 2007			
15.	Bielefeld	FU Berlin	Bonn	(#) → see text			

In this form, though, it is not easy to interpret, not least because each of these columns (A – E) is generated in a different manner.

"Research Strength" vs. Reputation

The "research strong" list (D) is generated by creating a research "profile" in which a point (colored green) is given if a program is in the T group in any one of the three indicators (e.g., found at rank 1-9 in A; rank 1-15 in B; or rank 1-12 in C). This is in absolute terms, and though it is not explained how, the positions are also assessed in relative terms. Thus, a program which ranks in the T groups on the three indicators in both absolute and relative terms can obtain a maximum of 6 points. Because a similar evaluation was conducted earlier,

the table also notes whether the program was "research strong" in 2004 (++) or it first joined in 2007 (+).

In a further differentiation, the CHE designates a university program as "research strong" if it has 4 or more of the 6 indicators. That status is reached by Giessen (6) and the LMU München (6), as well as by Freiburg (4), Münster (4) and Regensburg (4), though these last two are new members. If one ranks this way (by points, then by ++, then by +), one arrives at the order listed vertically in D. However, it is more accurate group them by their similarities and to give them ranks:

List D, ranks added

1. Giessen	3. Freiburg	4. Münster	6. Bayreuth	9. Konstanz
1. LMU München		4. Regensburg	6. FU Berlin	9. Mainz
			6. Heidelberg	
6++	4++	4+	3++	3+

It is worth noting already here that 3 universities (Münster, Bayreuth, Konstanz) on list D (indicator-based) do not appear on list E (reputation), and that 5 programs (HU Berlin, Tübingen, Frankfurt, Hamburg, Bremen) on list E (reputation) do not appear on list D (indicator-based). In other words, and most especially just below the top three programs, *reputation diverges from what the indicators reveal*.

The reputation list E is derived from asking professors in the discipline which universities they regard as "leading," and excluding mentioning one's own institution. Fot this list, the top group (LMU München, Freiburg, Giessen, and FU Berlin) is identified as those named by more than 25% of the respondents. If one takes the rank orders of lists D and E and compares them (see Table 2), one immediately sees that only 7 (of 16 total) programs appear on both:

Table 2

Comparing Rank Orders

Program	Rank List D	Rank List E	Actual % List E
LMU München	1	1	56
Giessen	1	3	40
Freiburg	3	2	53
Münster	4	--	--
Regensburg	4	10	7
Bayreuth	6	--	--

FU Berlin	6	4	31
Heidelberg	6	11	6
Konstanz	9	--	--
Mainz	9	8	8
HU Berlin	-	5	15
Tübingen	-	6	14
Frankfurt	-	7	11
Hamburg	-	9	8
Bremen	-	12	6
Köln	-	13	5

If one calculates r_s (Spearman rank-order correlation coefficient) just on these 7 programs that appear on both lists, one finds a respectable .63 correlation, suggesting that when data exists on both lists, there is moderately strong link between the collective indicators (D) and reputation (E). But when one does the same calculation for the entire list (substituting a rank of 10 for all missing values on the D list and a rank of 14 for missing values on the E list), this r_s value drops to .45, suggesting the connection is not nearly as strong when one value is missing. Put differently, information about high rank on the indicators, which is what the "research strong" designation provides, says little about reputation, and perhaps more telling, a good reputation says little about how "research strong" a program is.

In fact, one sees in miniature here a problem that bedevils many ranking efforts: it is easier to identify the very top than it is to determine what the right order is for ranks further down. If one calculates r_s without the top three (e.g., omitting Giessen, LMU München and Freiburg), for example, the correlation drops to a weak -.27, and if one takes the four pairs of the seven programs with values on both lists (and not in the top three), the correlation is only -.22. This exercise indicates that *while reputational measures can identify the very best programs, they are quite uncertain guides to the relative ranking of the rest*.

A different way to show this is to compare which programs are at ranks 1, 2 and 3 on the individual indicators and which are at ranks 4, 5, and 6:

Ranks	Funding	Publication	Ph.D. Completion
1.	LMU	Giessen	LMU
2.	Giessen	LMU	Heidelberg
3.	Freiburg	Freiburg	Freiburg

Here we have only 4 different universities (in 9 possible slots), and with one exception (at rank 11), they are those programs that by reputation are at ranks 1, 2 and 3. But just below these top three, we find that 'good' programs have different strengths:

Ranks	Funding	Publication	Ph.D. Completion
4.	HU Berlin	Mainz	Hamburg
5.	FU Berlin	Regensburg	Giessen
6.	Bremen	Frankfurt	Regensburg

This time we have 8 universities (in 9 possible slots), and they include programs that are at reputational ranks 3, 4, 5, 7, 8, 9, 10 and 12. The spread is much bigger, and it is impossible to say which single program deserves to be called fourth or fifth best, at least relative to these three indicators.

Yet another way to see this is to look at the third column in Table 2, and to note how weak reputational acknowledgement (meaning: which programs are called "leading") is for most programs. The top two programs (LMU, Freiburg) are clearly acknowledged as such by more than 50% of the respondents. But already the third-ranked (Giessen) is has only 2 of 5 respondents thinking well of it. By the fourth-ranked program (FU Berlin), we already have an anomaly: its reputation (at 31%) is vastly better (see the dicussion below) than the strength of its indicators.

Below rank 4, reputation rapidly declines. Three programs (HU Berlin, Tübingen, Frankfurt) have a respectable number of fans (10-15%), if few in number, but all remaining programs judged "leading"(Mainz & Hamburg (8%); Regensburg (7%); Heidelberg & Bremen (6%); Köln (5%) have very few of their colleagues thinking well of them. Given that about 50 programs are being evaluated, to be in this group of 13 programs which have a reputation at all as "leading" is doubtless an honor, though if one's program is toward the bottom (e.g. at 8% or less) it may be a case of damning with faint praise…

Indicators and Top (T), Middle (M) and Bottom (B) Groups

The indicator for the amount of outside funding spent (Table 1, column A), is derived from a CHE survey of *Anglistik* programs, and is updated by the programs themselves. While the sums cover a 3-year period (here, 2003-05), they are reported as absolute amounts by year and by professor. Outside funding includes resources from a wide variety of sources, with the

German Research Foundation (DFG) accounting for the largest single amount (41%), followed by the respective German state (11%), foundations (10%), the federal government (8%), the EU (7%), the German Academic Exchange Service (DAAD, also 7%). Private industry provides very little (3%), and there is also a residual "other" category (13%).

The 58 *Anglistik* programs overall reported spending about 5 million Euros; the cumulative addition down the list to 2.5 million spent (e.g., the top 50%) was reached by the 9th program. The bottom 1/5th (19 universities) of the list spent only 10%. On this indicator, T = 9, M = 21 and B = 19.

Functionally, this is an ordinal ranking, and not explicitly intended to be normative. Nevertheless, academic peers do hear about those programs that have more money, and if one sees in a ranking of this kind that the 1st-ranked program spent 100 times as much as the 49th-ranked, can one avoid thinking that *Anglistik* at the LMU München is simply **better** than at the Uni Trier? This line of thinking (they must be doing something (or are more clever) to bring in so much money they can spend) is what leads to efforts to try to imitate that success (e.g., to engage in 'benchmarking,' one way one might interpret the *Exzellenzinitiative*). The distribution here appears to follow a power law: the top two programs spent 1.5x as much as the next two programs, and those two in turn spent twice as much as the fifth-ranking program. The top program spent 3.3 times as much as the 10th program.

One can also look at this from the input side. The DFG publishes information on the total volume of DFG awards it makes, by institution and by broad academic category. If one calculates the rank order correlation (not shown here) between the 13 *Anglistik* programs regarded as "leading" with information on the amount of DFG funding (absolute, 2005-07) awarded to the humanities and social sciences in these 13 institutions, one finds $r_s = -.52$, which is a relatively strong link. One might suggest (but only suggest!) that part of the reputation for having a "leading" *Anglistik* program derives from a more general impression that a particular university is more successful at bringing in funding (in this case, from the major single outside source of funding for *Anglistik*) for humanities to the institution. This halo effect could explain why certain *Anglistik* programs (such as at the FU Berlin or the HU Berlin) have reputations that do not seem to match their positions on the CHE indicators.

The third indicator, Ph.D. completion rates (column C) is also derived from the survey of programs. In this case it refers to the six semesters prior to the questionnaire (summer 2003 to the winter of 2005-6), and it is also reported by year and professor. Here, the numbers by group are T = 12, M = 26 and B = 13. The total numbers of graduates is very small (there are five programs in the list, for example, that only graduated a single Ph.D. over

the three-year period), and it follows a power-law distribution. The top program (at 15.3 per year) had more than twice as many as the second (6.7 per year), and the long tail of the distribution includes 22 (of 51) programs that graduated less than 2 Ph.D.s per year. The top program graduated 4.6 times as many Ph.D.s as the 10th program.

If one then combines the rankings for D and E, and adds information about where, on the three indicators, the individual programs fall in terms of Top, Middle, and Bottom positions in the overall list of all programs (see Table 3a), one finds some interesting patterns that can help in interpreting the reputational rankings. The Table here is ordered by the descending number of times a program is found in the top groups. Within each of these sub-groups, M >B (as in T = 2), and then higher position on indicator B (as in T =3) determines the order. Actual position is given for all T values; bold superscripts indicate the top rank on that indicator, underlined subscripts indicate the bottom rank in that T (or M) group.

Table 3a

Programs by relative position on the indicators

Top	Rank List D	Rank List E	% E	Top/Middle/Bottom Groups A	B	C
T = 3						
LMU	1	1	56	$T^{\mathbf{(1)}}$	$T_{(2)}$	$T^{\mathbf{(1)}}$
Giessen	1	3	40	$T_{(2)}$	$T^{\mathbf{(1)}}$	$T_{(7)}$
Freiburg	3	2	53	$T_{(\underline{9})}$	$T_{(3)}$	$T_{(2)}$
Münster	4	--	--	$T_{(5)}$	$T_{(8)}$	$T_{(\underline{12})}$
T = 2						
Frankfurt	--	7	11	M	$T_{(7)}$	$T_{(11)}$
Heidelberg	6	11	6	M	$T_{(\underline{9})}$	$T_{(3)}$
Mainz	9	8	8	B	$T_{(5)}$	$T_{(10)}$
Regensburg	4	10	7	B	$T_{(6)}$	$T_{(8)}$
T = 1						
Köln	--	13	5	M	$T_{(11)}$	M
Konstanz	9	--	--	M	$T_{(13)}$	M
FU Berlin	6	4	31	M	$T_{(\underline{15})}$	M
Hamburg	--	9	8	M	M	$T_{(6)}$
Bayreuth	6	--	--	$T_{(8)}$	$M_{(\underline{38})}$	M
T = 0						
Bremen	--	12	6	M	M	M
HU Berlin	--	5	15	M	B	M
Tübingen	--	6	14	--	M	--

If one reorders this table (see Table 3b) by actual percentages received (along the the %E column), one can see just how problematic reputation is as a measure. There are programs with fairly strong indicators (say, defined as being in the T group twice) whose reputation does not match it (esp. Münster, but also Frankfurt, Mainz, Regensburg, and Heidelberg). There are also programs with relatively high reputations (FU Berlin, HU Berlin, Tübingen) which must have them for reasons other than what is reflected by these indicators (perhaps their reputation reflects past glory?), as most of them are only in the middle range. One can also note the relative *un*importance of the financial indicator (A) – in this list of 16 programs, only 5 in the top group – for reputation as compared with publication (11 of 16), the topic we turn to next.

Table 3b

Programs ordered by actual reputational percentages

Top	Rank List D	Rank List E	% E	Top/Middle/Bottom Groups A	B	C	
							Rank 1-3
LMU	1	1	56	$T^{(1)}$	$T_{(2)}$	$T^{(1)}$	T = 9
Freiburg	3	2	53	$T_{(9)}$	$T_{(3)}$	$T_{(2)}$	M = 0
Giessen	1	3	40	$T_{(2)}$	$T^{(1)}$	$T_{(7)}$	B = 0
							Rank 4-7
FU Berlin	6	4	31	M	$T_{(15)}$	M	T = 3
HU Berlin	--	5	15	M	B	M	M = 6
Tübingen	--	6	14	--	M	--	B = 1
Frankfurt	--	7	11	M	$T_{(7)}$	$T_{(11)}$	Missing = 2
							Rank 8-13
Mainz	9	8	8	B	$T_{(5)}$	$T_{(10)}$	T = 8
Hamburg	--	9	8	M	M	$T_{(6)}$	M = 8
Regensburg	4	10	7	B	$T_{(6)}$	$T_{(8)}$	B = 2
Heidelberg	6	11	6	M	$T_{(9)}$	$T_{(3)}$	
Bremen	--	12	6	M	M	M	
Köln	--	13	5	M	$T_{(11)}$	M	
							Unranked
Münster	4	--	--	$T_{(5)}$	$T_{(8)}$	$T_{(12)}$	T = 5
Konstanz	9	--	--	M	$T_{(13)}$	M	M = 4
Bayreuth	6	--	--	$T_{(8)}$	$M_{(38)}$	M	B = 0
Cumulative No. of T ranks				5	11	9	

The Publication Indicator

Though the CHE called it a "bibliometric" survey, it did not conduct an analysis based on citation data but instead collected biblio*graphic* data and weighted it. The basis was provided by the publications of professors and postdoc researchers (*promovierte Wissenschaftler*) from 2003-2005. It included all those employed in the department, whether on stipends, budget lines, or outside funding. The numbers are based on current faculty, the names corrected and augmented by the administrators in the respective institutions. If an individual changed institutions, publications in the 2003-05 period were ascribed to the current university even if the work was done at a previous institution. In other words, this indicator is based not on institution but on persons producing the publications, even though the indicator itself only lists institutions.

The basis for the publication data is the "Annual Report on English and American Studies," which compiles a yearly biography. Included in it are monographs, articles that can be assigned to journals in the field (an interesting definition!), contributions to edited works, and encyclopaedia articles. Publications were weighted by length, such that those up to 4 pages received 1 point, 5-9 pages 2 points, 10-19 pages 3 points, 20-39 4 points, 40-99 5 points and 100+ 8 points (for collaborations, points were awarded proportionate to the number of, so for two authors they were halved, for three in thirds, and so forth). Editorships, even if multiple, were given 2 points. The exact formula used for these calculations were not provided, nor were examples given.

The analysis looked not only at the absolute number of publications, but also calculated a 'publication per professor' rate. As with the other indicators, the 53 programs were divided into top (15), middle (23), and bottom (15) groups. Table 4 provides the data:

Table 4

Publication Indicator, by group

	Number of universities	Percent of total pubs.	Total number publications (range per university)	Mean
Top group	15	50	1410 (183 – 72)	94 (88, corrected)
Middle group	23	40	1023 (69 – 28)	44
Bottom group	15	10	298 (28 – 4)	20

The first three ranks of the top group, had 183 (Giessen), 124 (LMU München), and 118 (Freiburg) publications, respectively. That means Giessen published as much as the bottom 10 programs combined, and had 50% more publications than the LMU München, the second-ranked. This has the effect of skewing the average in the top group up, so the corrected average (omitting Giessen) is also calculated and given.

But the differences between the means of the groups nevertheless permit a striking conclusion: the top *Anglistik* departments (in publications), on average publish **twice** as much as the middle group, and the middle group, in turn, publishes **twice** as much as the programs in the bottom group. Quantity does indeed matter, at least in this CHE comparison.

CHE also provides a more complex measure, namely publications per professor. A frequency count indicates the following:

Publication per Professor	Number of programs
<1	1
1.0-1.9	4
2.1-2.9	7
3.1-3.9	10
4.1-4.9	14
5.1-5.9	10
6.1-6.9	0
7.1-7.9	3
8.1-8.9	3
9.0+	1

This is a near normal distribution, with the mode at 4-5, and an interesting gap near the top of the frequency distribution. The averages by group permit a similar observation as before:

	Mean publication points
Top group (15)	**6.07** (without Giessen: 5.57)
Middle group (23)	**4.37**
Bottom group (15)	**2.88** (without Leipzig: 2.45)

the top, middle and bottom groups can be identified by the number of publications, when measured on a per professor basis.

Nevertheless, publication does not reputation make. Or rather, if one takes the top publications by professor (which may a more accurate reflection of productivity), a number of programs appear (Bamberg, Osnabrück, Duisburg, Eichstatt) which otherwise are not much in evidence – and again, there are programs on the reputational list (esp. the HU Berlin, though Hamburg to a lesser extent) whose overall publication amount and publication by professor

rates are simply very low. Only four programs (Giessen, LMU, FU Berlin and Regensburg) which have high publications per professor rates (e.g., putting them in the top group) also appear on the reputational ranking list.

Concluding Reflections

It should be clear from this small exercise that while both peer review and an indicator-based ranking of programs may have their respective merits, they yield different results – and in that sense cannot be treated as two ways to arrive at the same judgment. Or rather, at the very top they yield essentially the same result, but as soon as one leaves that level, one finds only a moderate correlation, whether it is a particular indicator (including total number of publications) with reputation or reputation with a particular indicator. There appear to be programs whose reputations as "leading" are not supported by the indicators examined here (cf. HU Berlin and publication), and there are programs whose indicators are quite strong (cf. Münster) without this translating into a reputation for having a "leading" program.

As it is, these may not be terribly helpful indicators for this particular discipline. Third-party funding does not play nearly the role in the humanities that it does in other fields of academic endeavour, and it may also simply not be very meaningful to differentiate between an Uni Köln (at rank 26, spending 56,000, M group) and an RWTH Aachen (at rank 30, spending 50,000, B group). Arguments about the use of third-party funding at all as an indicator have also been made (cf. Fraunhofer study).

The same can be said even more strongly of using Ph.D. graduation rates as an indicator. It's not entirely clear what this measures – good vs. bad advising? extending pleasant student life a few more years? the availability of part-time employment vs. the economic pressures to get onto the labor market quickly with a degree in hand? This may be a case of counting it simply because it can be measured. This CHE survey also shows a very small number of such students graduating; as above, many fine distinctions may be getting drawn here that show differences between programs that do not, in practice, exist.

Obviously, the point here is to identify the "very best" programs, and that both reputational rankings *and* these indicators can do. If one wants to genuinely understand what is going on within programs, what their strengths and weakness are, and what their future potential might be, then one has to engage in the kind of analysis of *Anglistik* the Land Niedersachsen carried out in 2004, and more closely examine each program in turn. That the

evaluation of research per se in that particular study ended up being only one factor of many raises a question of how useful such a tool is – other than to tell state education ministries something about the policy choices they face.

The bottom line from examining this CHE report, however, is that neither peer review nor indicators do a particularly good job. The worth of a program, as judged by the indicators, may well be ignored by the peers. And the peers may judge a program to be good, in evident ignorance of how little it is actually producing by way of publications or graduates. Whether either side could properly learn from the other is an open question.

References :

Berghoff, Sonja et. al. *Das CHE-ForschungsRanking deutscher Universitäten 2007*, Arbeitspapier Nr. 102 (Februar 2008). Available at: CHE-ForschungsRanking_2007_AP_102.pdf or at http://www.academics.de/image-upload/CHE_ForschungsRanking_Anglistik_2007_0.pdf

Wissenschaftliche Kommission Niedersachsen. 2004. Forschungsevaluation an Niedersächsichen Hochschulen und Forschungseinrichtungen: Anglistik und Amerikanisti. Ergebnisse und Empfehlungen.